System Identification
Lecture 7
PE Method - Variance and Uncertainty Intervals

Paul Van den Hof

Control Systems Group
Department of Electrical Engineering
Eindhoven University of Technology

Contents

Lecture 7: Variance and uncertainty intervals

Contents Chapter 4 - Part 4

- (Asymptotic) distribution and variance of estimated parameters
- Parameter uncertainty regions
- Variance and uncertainty regions of model's frequency response
- Minimum variance estimator and Cramer-Rao Lower Bound (CRLB)

Statistical properties of $\hat{\theta}_N$ when $\mathcal{S} \in \mathcal{M}$

Due to the stochastic noise $v(t)$ corrupting the data Z^N, the identified parameter vector $\hat{\theta}_N$ is a random variable i.e.

the value of $\hat{\theta}_N$ is different at each experiment

Illustration of the statistical properties of $\hat{\theta}_N$

$$\mathcal{S}: \quad y(t) = \frac{0.7q^{-1}}{1+0.3q^{-1}}u(t) + \frac{1}{1+0.3q^{-1}}e(t) \quad (\sigma_e^2 = 0.3)$$

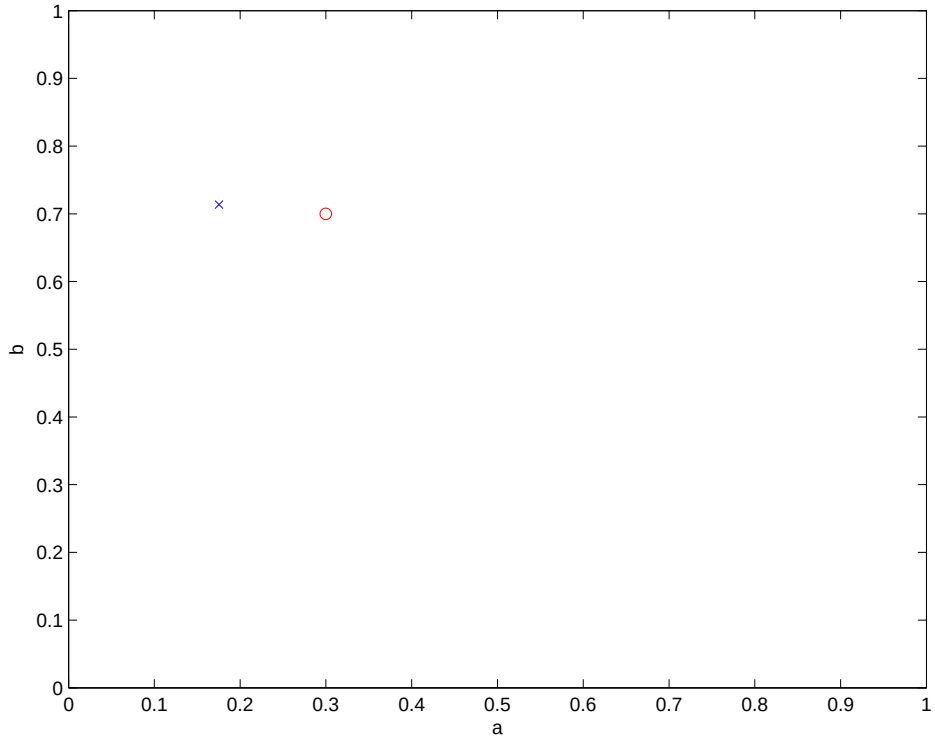
$$\mathcal{M}: \quad G(q,\theta) = \frac{bq^{-1}}{1+aq^{-1}} \quad H(q,\theta) = \frac{1}{1+aq^{-1}} \quad \theta = \begin{pmatrix} a \\ b \end{pmatrix}$$

we have applied a white noise $u(t)$ of length $N = 200$ and of power $\sigma_u^2 = 0.5$

we have measured the corresponding $y(t)$.

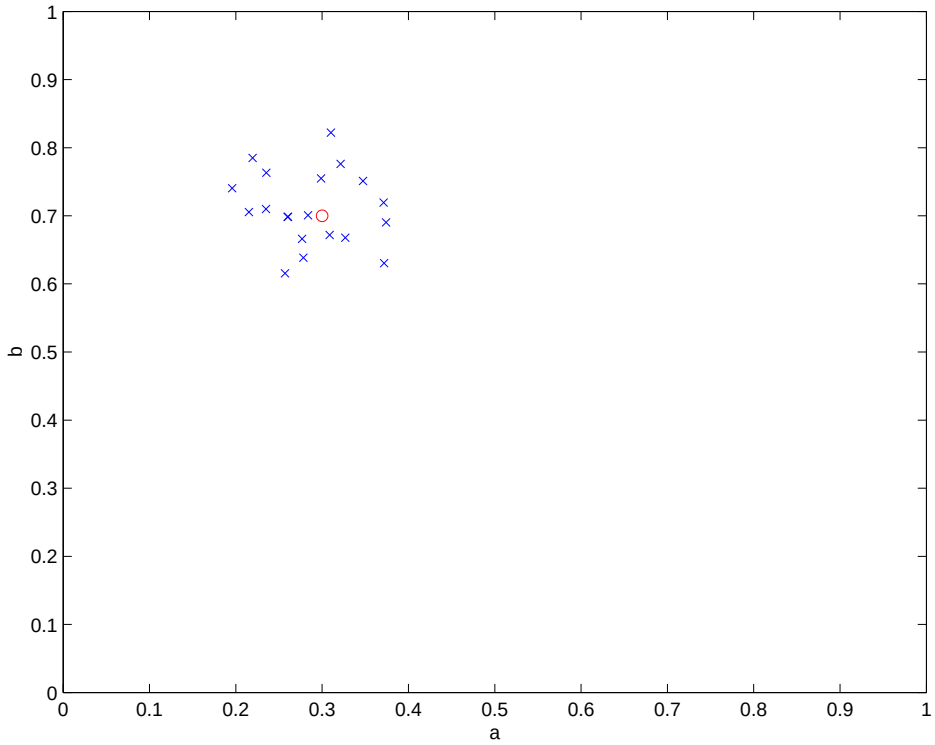
we have computed the estimate $\hat{\theta}_N$ of $\theta_0 = \begin{pmatrix} 0.3 & , & 0.7 \end{pmatrix}^T$

The estimate $\hat{\theta}_N$ is represented with a blue cross and θ_0 by a red circle.

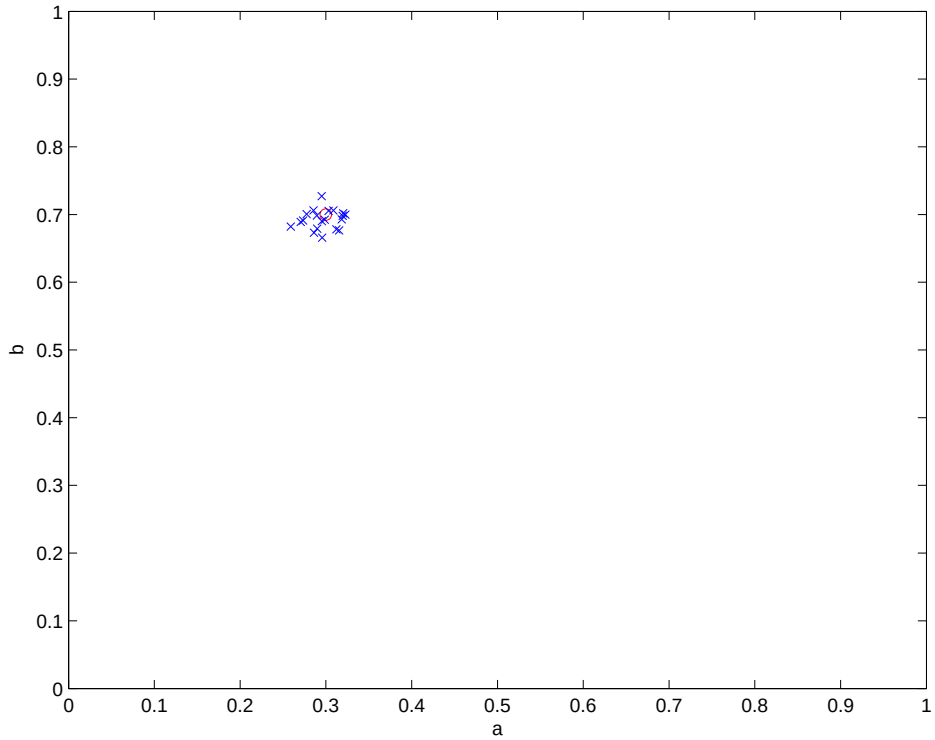


Repetition of the experiment
we have applied 20 times the same sequence $u(t)$ of length $N = 200$ and power $\sigma_u^2 = 0.5$.
every experiment has a different realization of $v(t)$
we have measured the corresponding $y(t)$
For these 20 experiments, we have computed the estimate $\hat{\theta}_N$

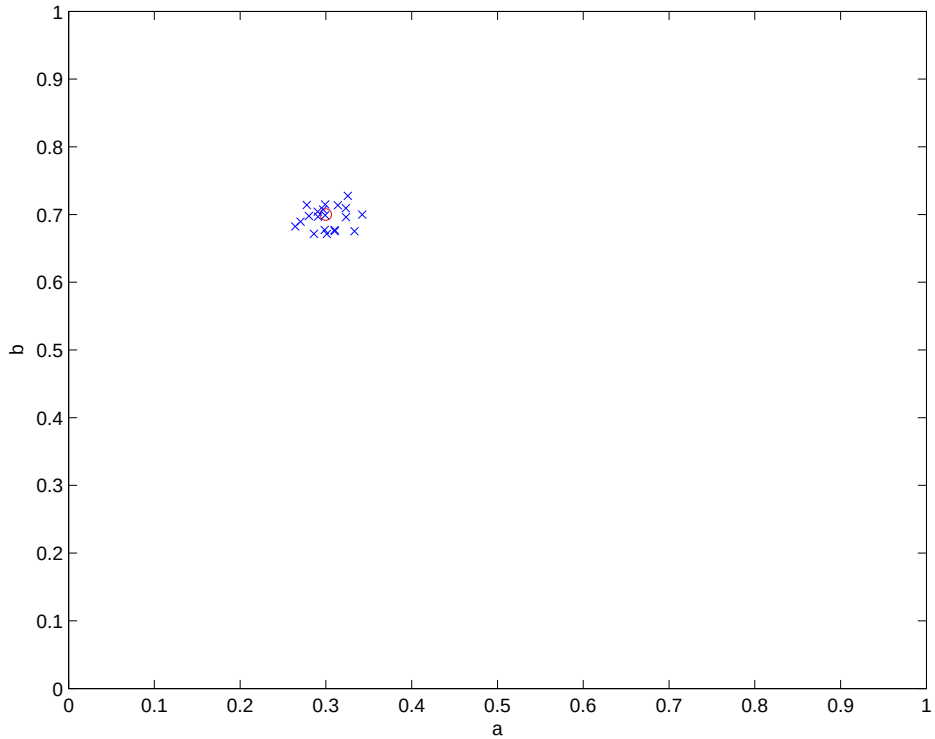
The twenty estimates $\hat{\theta}_N$ are represented with a blue cross and θ_0 by a red circle.



Let us now change the experimental conditions. We have applied 20 times the same sequence $u(t)$ of length $N = 2000$ and power $\sigma_u^2 = 0.5$. The twenty estimates $\hat{\theta}_N$ are represented with a blue cross and θ_0 by a red circle.



Another change of experimental conditions. We have applied 20 times the same sequence $u(t)$ of length $N = 200$ and power $\sigma_u^2 = 5$. The twenty estimates $\hat{\theta}_N$ are represented with a blue cross and θ_0 by a red circle.



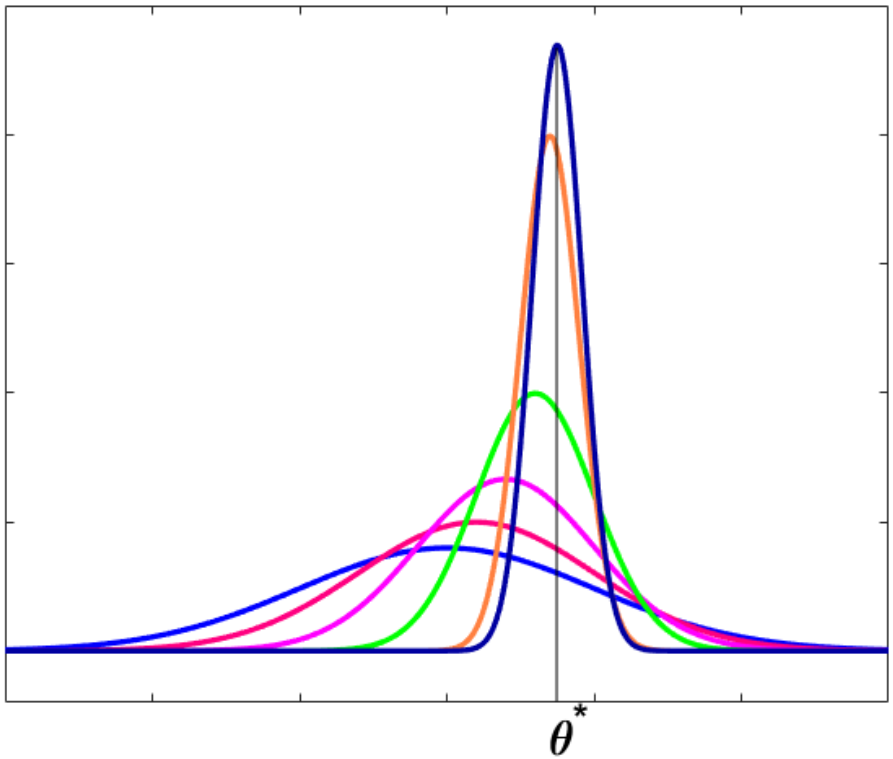
Observations:

- θ_0 seems to be the mean of $\hat{\theta}_N$
- the variance seems to decrease when N and/or the power of u are larger

These observations are confirmed by the properties of the estimate $\hat{\theta}_N$ presented in the sequel

Asymptotic distribution

Parameter estimate $\hat{\theta}_N$ has a pdf that is dependent on N:



The asymptotic pdf of the estimate $\hat{\theta}_N$ is specified as follows:

Asymptotic distribution

Consider the data assumptions as present in the convergence result. Then for $N \to \infty$,

$$\sqrt{N}(\hat{\theta}_N - \theta^*) \to \mathcal{N}(0, P_\theta)$$

where, in the situation $\theta^* = \theta_0$ ($\mathcal{S} \in \mathcal{M}$), the covariance matrix P_θ is given by

$$P_\theta = \sigma_e^2 \cdot \left[\bar{\mathbb{E}}\psi(t, \theta_0)\psi^T(t, \theta_0)\right]^{-1}$$

$$\psi(t, \theta_0) := \left.\frac{\partial \hat{y}(t|t-1;\theta)}{\partial \theta}\right|_{\theta=\theta_0} = -\left.\frac{\partial \varepsilon(t, \theta)}{\partial \theta}\right|_{\theta=\theta_0}$$

Justification of normal distribution

As example: linear regression structure, $\mathcal{S} \in \mathcal{M}$, leading to

$$\hat{\theta}_N - \theta_0 = \left[\frac{1}{N}\sum_t \varphi(t)\varphi^T(t)\right]^{-1} \cdot \frac{1}{N}\sum_t \varphi(t)e(t)$$

with $e(t)$ a random variable and e a white noise process.
Then, for $N \to \infty$, the above expression becomes a weighted sum of an infinite number of random variables with a fixed distribution. According to the law of large numbers, this weighted sum will get a Gaussian pdf.

P_θ/N has the interpretation of covariance matrix of the parameter estimate:

$$P_\theta/N = \mathbb{E}[(\hat{\theta}_N - \theta_0)(\hat{\theta}_N - \theta_0)^T].$$

quantifying the variability of the estimate

Covariance matrix:

Let $\tilde{\theta} := \hat{\theta}_N - \theta_0$, then

$$P_\theta/N = \mathbb{E}\begin{bmatrix} \tilde{\theta}_1 \\ \tilde{\theta}_2 \\ \vdots \\ \tilde{\theta}_n \end{bmatrix} \cdot \begin{bmatrix} \tilde{\theta}_1 & \tilde{\theta}_2 & \cdots & \tilde{\theta}_n \end{bmatrix} \quad \in \mathbb{R}^{n\times n}$$

and

$$[P_\theta/N]_{jk} = \mathbb{E}\tilde{\theta}_j\tilde{\theta}_k$$

Parameter variances are on the diagonal of P_θ/N.

Justification of the covariance matrix expression

In the linear regression situation, $\mathcal{S} \in \mathcal{M}$, we have

$$\hat{\theta}_N - \theta_0 = \left[\frac{1}{N}\sum_t \varphi(t)\varphi^T(t)\right]^{-1} \cdot \frac{1}{N}\sum_t \varphi(t)e(t)$$

so that $\bar{\mathbb{E}}\tilde{\theta}\tilde{\theta}^T$ can be written as

$$\bar{\mathbb{E}}\left\{\left[\frac{1}{N}\sum_t \varphi(t)\varphi^T(t)\right]^{-1} \frac{1}{N}\sum_t \varphi(t)e(t)\cdot \right.$$
$$\left. \cdot\frac{1}{N}\sum_t \varphi^T(t)e(t)\left[\frac{1}{N}\sum_t \varphi(t)\varphi^T(t)\right]^{-1}\right\}$$

$P_\theta/N = E(\hat{\theta}_N - \theta_0)(\hat{\theta}_N - \theta_0)^T$ can be estimated from the data and $\hat{\theta}_N$ as:

$$\begin{aligned} \hat{P}_\theta &= \hat{\sigma}_e^2 [\frac{1}{N} \sum_{t=1}^{N} \psi(t, \hat{\theta}_N) \psi^T(t, \hat{\theta}_N)]^{-1} \\ \hat{\sigma}_e^2 &= \frac{1}{N} \sum_{t=1}^{N} \varepsilon^2(t, \hat{\theta}_N) \end{aligned}$$

with $\hat{\sigma}_e^2$ an estimate of σ_e^2.

Note:

The asymptotic distribution result makes a statement about the shape of the pdf of $\hat{\theta}_N$ for $N \to \infty$

The consistency result states that for $N \to \infty$ the pdf of $\hat{\theta}_N$ will become a Dirac pulse at $\hat{\theta}_N = \theta_0$.

Therefore

- the covariance matrix $P_\theta/N \to 0$ when $N \to \infty$
- P_θ/N tells us how fast the convergence to the Dirac pulse appears.

Some observations on the covariance matrix P_θ/N of $\hat{\theta}_N$

$$\frac{P_\theta}{N} \approx \frac{\sigma_e^2}{N}\left(\frac{1}{N}\sum_{t=1}^{N}\psi(t,\hat{\theta}_N)\psi^T(t,\hat{\theta}_N)\right)^{-1}$$

Observation 1. The larger N, the smaller P_θ/N

Observation 2. The larger the power of $u(t)$, the smaller P_θ

The value of P_θ/N can be influenced by an appropriate choice of the experimental conditions

"Proof" of observation 2: P_θ is proportional to the inverse of the power of the vector signal $\psi(t, \hat{\theta}_N)$ and this vector signal has higher power when $u(t)$ has higher power (when focusing on the parameters in G). Indeed

$$\begin{aligned} \epsilon(t,\theta) &= H^{-1}(q,\theta)\left(y(t) - G(q,\theta)u(t)\right) \\ &= \frac{G_0(q) - G(q,\theta)}{H(q,\theta)}u(t) + \frac{H_0}{H(q,\theta)}e(t) \end{aligned}$$

and $\psi(t,\theta) = -\left.\frac{\partial \varepsilon(t,\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}_N}$

Parameter uncertainty regions

The asymptotical normal distribution can be used to quantify parameter uncertainty regions.
Start with:

$$\sqrt{N}(\hat{\theta}_N - \theta_0) \to \mathcal{N}(0, P_\theta)$$

Let P_θ^{-1} be decomposed as $P_\theta^{-1} = R_\theta^T R_\theta$, then

$$\sqrt{N} R_\theta(\hat{\theta}_N - \theta_0) \to \mathcal{N}(0, I_n)$$

i.e. a vector of n standard (independent) normally distributed random variables.

If $x \in \mathcal{N}(0, I_n)$ then for the sum of the quadratic variables:

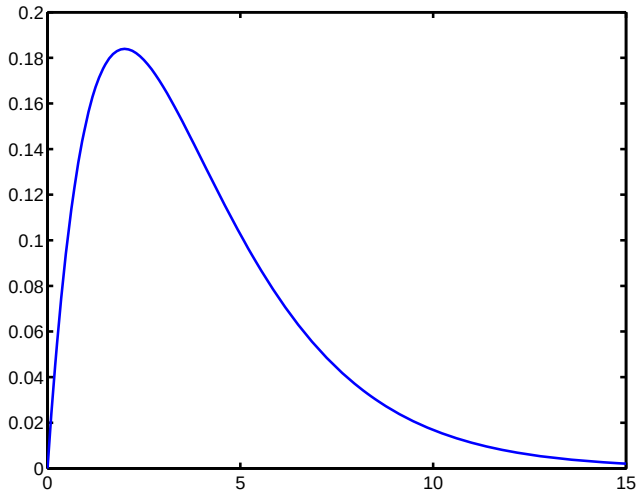
$$x^T x \in \chi^2(n)$$

i.e. follows a χ^2 distribution, scalar valued.

This implies that the sum of quadratic variables

$$N(\hat{\theta}_N - \theta_0)^T P_\theta^{-1}(\hat{\theta}_N - \theta_0) \to \chi^2(n).$$

Example χ^2-distribution



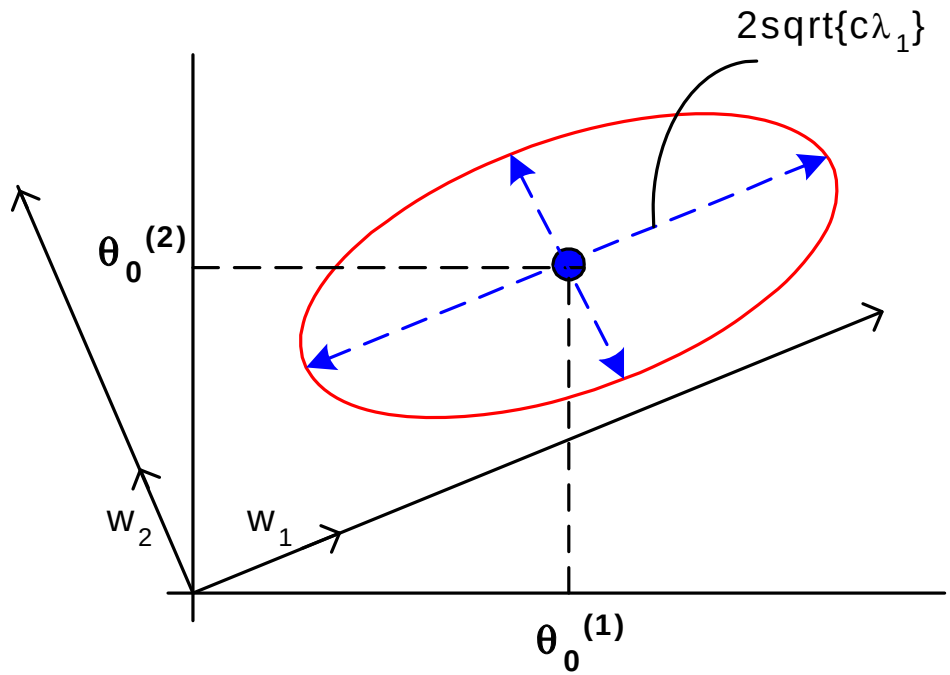
For $\mathbf{x} \in \chi^2(n)$:

$$\int_0^{c_\chi(\alpha,n)} x \; dx = \alpha$$

Note that contour lines:

$$N(\hat{\theta}_N - \theta_0)^T P_\theta^{-1}(\hat{\theta}_N - \theta_0) = c$$

are ellipsoidal contour lines of the multivariable Gaussian distribution:



(P_θ has eigenvalues λ_i and eigenvectors w_i)

while the level of probability related to the event

$$N(\theta - \theta_0)^T P_\theta^{-1}(\theta - \theta_0) < c$$

is determined by the $\chi^2(n)$-distribution.

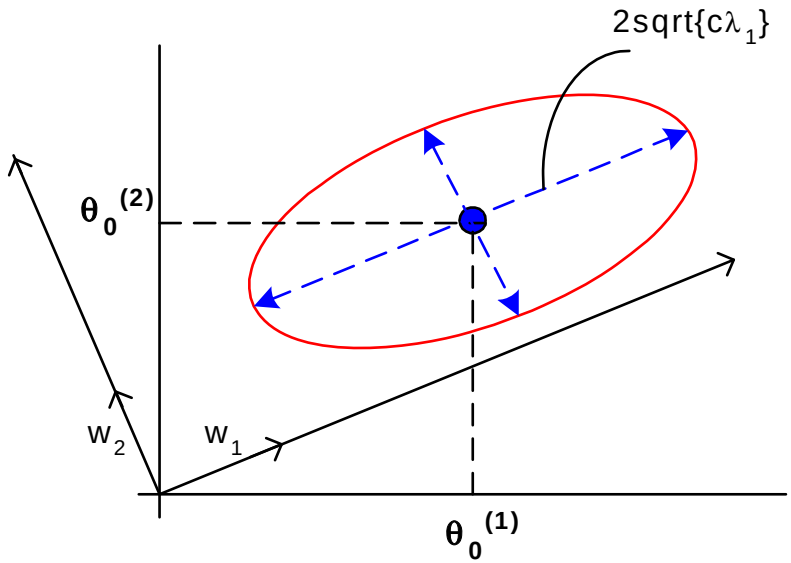
Denote with $c_\chi(\alpha, n)$ the value that satisfies

$$Pr(x \leq c_\chi(\alpha, n)) = \alpha$$

for $x \in \chi^2(n)$.
Then $\hat{\theta}_N \in \mathcal{D}_{\theta_0}$ with probability α, with

$$\mathcal{D}_{\theta_0} = \{\theta \mid (\theta - \theta_0)^T P_\theta^{-1} (\theta - \theta_0) < \frac{c_\chi(\alpha, n)}{N}\}$$

ellipsoid centered at θ_0.



Alternatively, denote:

$$\mathcal{D}_{\hat{\theta}_N} = \{\theta \mid (\theta - \hat{\theta}_N)^T P_\theta^{-1} (\theta - \hat{\theta}_N) < \frac{c_\chi(\alpha, n)}{N}\}$$

ellipsoid centered at $\hat{\theta}_N$
then it can simply be verified that

$$\hat{\theta}_N \in \mathcal{D}_{\theta_0} \Leftrightarrow \theta_0 \in \mathcal{D}_{\hat{\theta}_N}$$

so that $\theta_0 \in \mathcal{D}_{\hat{\theta}_N}$ with probability α.

Example:

$$\mathcal{S}: \quad y(t) = \frac{0.7q^{-1}}{1+0.3q^{-1}}u(t) + \frac{1}{1+0.3q^{-1}}e(t)$$

$$\mathcal{M}: \quad G(q,\theta) = \frac{bq^{-1}}{1+aq^{-1}}; \quad H(q,\theta) = \frac{1}{1+aq^{-1}}; \quad \theta = \begin{pmatrix} a \\ b \end{pmatrix}$$

we have applied a sequence $u(t)$ of length $N = 1000$ to $\mathcal{S}$ and we have measured the corresponding $y(t)$.

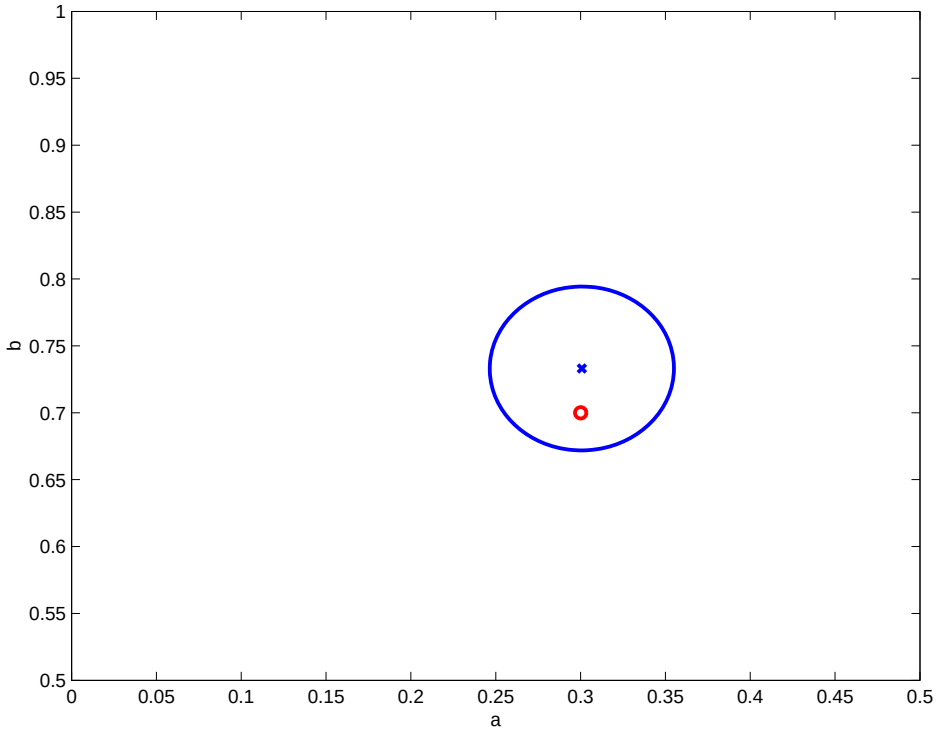
Using these data, we have computed the estimate $\hat{\theta}_N$ of $\theta_0 = \begin{pmatrix} 0.3 & , & 0.7 \end{pmatrix}^T$ along with its (estimated) covariance matrix P_θ:

$$\hat{\theta}_N = \begin{pmatrix} 0.301 \\ 0.733 \end{pmatrix} \qquad \hat{P}_\theta = \begin{pmatrix} 0.4922 & 0.0017 \\ 0.0017 & 0.6264 \end{pmatrix}$$

The 95% uncertainty region $\mathcal{D}$ can then be constructed

$$\mathcal{D} = \left\{ \theta \in \mathbb{R}^2 \mid (\theta - \hat{\theta}_N)^T \ \hat{P}_\theta^{-1} \ (\theta - \hat{\theta}_N) \ \leq 5.99 \cdot 10^{-3} \right\}$$

The estimate $\hat{\theta}_N$ (blue cross) along with its uncertainty ellipsoid $\mathcal{D}$ in the parameter space



The in practice unknown θ_0 is represented by the red circle and lies in $\mathcal{D}$ as expected

Parameter uncertainties for each parameter separately.

Based on the ellipsoid, an uncertainty region for each separate parameter can also be specified.

From $\hat{\theta}_N \in \mathcal{N}(\theta_0, P_\theta/N)$ it follows that for parameter $\hat{\theta}_N^{(i)}$ it holds that

$$\hat{\theta}_N^{(i)} \in \mathcal{N}(\theta_0^{(i)}, P_\theta^{(ii)}/N)$$

with $P_\theta^{(ii)}$ the i-th diagonal element of P_θ.

This leads to the choice of an uncertainty region:

$$|\hat{\theta}_N^{(i)} - \theta_0^{(i)}| < 3\sqrt{P_\theta^{(ii)}/N}$$

that is satisfies with a probability level of 99% (3 σ-bound).

Uncertainty regions for frequency responses of estimated models

The identified parameter vector $\hat{\theta}_N$ is a random variable distributed as $\hat{\theta}_N \sim As\mathcal{N}(\theta_0, P_\theta/N) \Longrightarrow$

the identified models (frequency responses) $G(e^{i\omega}, \hat{\theta}_N)$ (and $H(e^{i\omega}, \hat{\theta}_N)$) are also random variables:

- $G(e^{i\omega}, \hat{\theta}_N)$ is an (asymptotically) unbiased estimate of $G(e^{i\omega}, \theta_0)$
- the variance of $G(e^{i\omega}, \hat{\theta}_N)$ is defined in the frequency domain as:

$$cov(G(e^{i\omega}, \hat{\theta}_N)) := \mathbb{E}\left(|G(e^{i\omega}, \hat{\theta}_N) - G(e^{i\omega}, \theta_0)|^2\right)$$

Covariance of the frequency response

$$cov\{G(e^{i\omega}, \hat{\theta}_N\} = \mathbb{E}\left(|G(e^{i\omega}, \hat{\theta}_N) - G(e^{i\omega}, \theta_0)|^2\right)$$

is obtained by

$$cov\{G(e^{i\omega}, \hat{\theta}_N\} \sim \left.\frac{\partial G(e^{i\omega}, \theta)}{\partial \theta}^*\right|_{\theta_0} \cdot \frac{P_\theta}{N} \cdot \left.\frac{\partial G(e^{i\omega}, \theta)}{\partial \theta}\right|_{\theta_0}$$

which is a first order Taylor approximation that is exact for models that are linear in the parameters.
$((\cdot)^*$ is complex conjugate transpose)
$cov\{G(e^{i\omega}, \hat{\theta}_N\}$ can be estimated by replacing P_θ by $\hat{P}_\theta$, and θ_0 by $\hat{\theta}_N$.

Proof of the expression of $cov(G(e^{i\omega}, \hat{\theta}_N))$

First order (Taylor) approximation:

$$G(e^{i\omega}, \hat{\theta}_N) \approx G(e^{i\omega}, \theta_0) + \left(\hat{\theta}_N - \theta_0\right)^T \Lambda_G(e^{i\omega}, \theta_0)$$

with $\Lambda(e^{i\omega}, \theta_0) = \left.\frac{\partial G(e^{i\omega},\theta)}{\partial \theta}\right|_{\theta_0}$.

Consequently: $|G(e^{i\omega}, \hat{\theta}_N) - G(e^{i\omega}, \theta_0)|^2 = ..$

$$\begin{aligned} ... &= (G(e^{i\omega}, \hat{\theta}_N) - G(e^{i\omega}, \theta_0))^*(G(e^{i\omega}, \hat{\theta}_N) - G(e^{i\omega}, \theta_0)) \\ &\approx \Lambda_G(e^{i\omega}, \theta_0)^* \left(\hat{\theta}_N - \theta_0\right)\left(\hat{\theta}_N - \theta_0\right)^T \Lambda_G(e^{i\omega}, \theta_0) \end{aligned}$$

$$E(.) = \Lambda_G(e^{i\omega}, \theta_0)^* \cdot \frac{P_\theta}{N} \cdot \Lambda_G(e^{i\omega}, \theta_0)$$

Observation

The larger N and/or the larger the power of $u(t)$, the smaller $cov(G(e^{i\omega}, \hat{\theta}_N))$

direct consequence of the fact that P_θ/N has this property

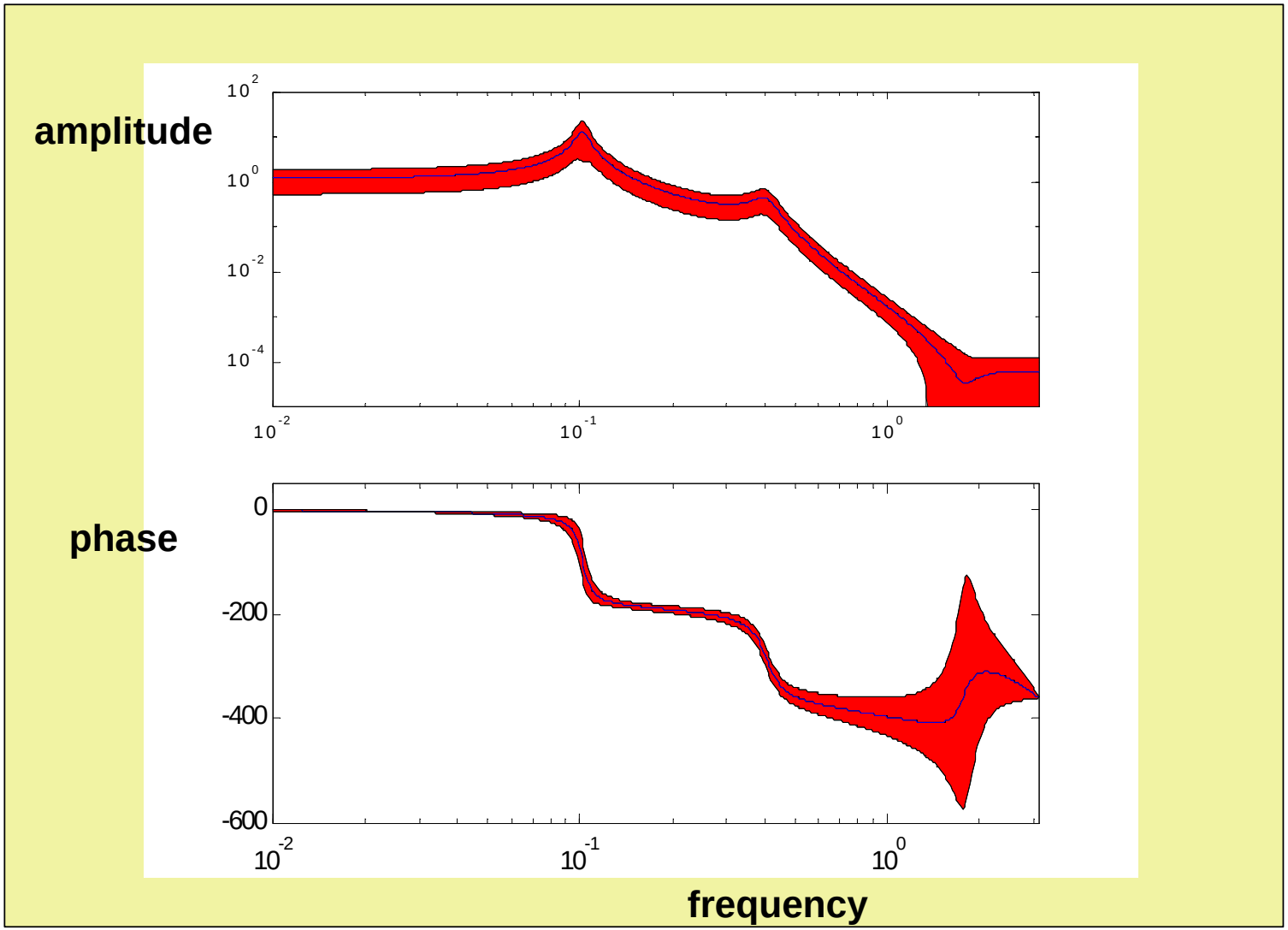
Separate bounds for amplitude and phase of $G(e^{i\omega}, \hat{\theta}_N)$

$$
\begin{aligned}
f_{a,\omega}(\theta) &= |G(e^{i\omega}, \theta)|, \\
f_{p,\omega}(\theta) &= arg\{G(e^{i\omega}, \theta)\}
\end{aligned}
$$

then covariance information on $f_{a,\omega}$ and $f_{p,\omega}$ is obtained from the first order approximations

$$
\begin{aligned}
Cov\{f_{a,\omega}(\hat{\theta}_N)\} &\approx \left.\frac{\partial f_{a,\omega}(\theta)}{\partial \theta}^T\right|_{\theta=\hat{\theta}_N} \frac{P_\theta}{N} \left.\frac{\partial f_{a,\omega}(\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}_N} \\
Cov\{f_{p,\omega}(\hat{\theta}_N)\} &\approx \left.\frac{\partial f_{p,\omega}(\theta)}{\partial \theta}^T\right|_{\theta=\hat{\theta}_N} \frac{P_\theta}{N} \left.\frac{\partial f_{p,\omega}(\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}_N}
\end{aligned}
$$

This is implemented in Matlab's System Identification Toolbox (3σ bounds).


amplitude
phase
frequency
10^2
10^0
10^-2
10^-4
10^-2
10^-1
10^0
0
-200
-400
-600

Remarks

- Estimated uncertainty bounds are reliable only if the models are correct (validated), i.e. $\mathcal{S} \in \mathcal{M}$, or $G_0 \in \mathcal{G}$.
- If the uncertainty bound for $G(e^{i\omega}, \hat{\theta}_N)$ is too large in a particular frequency range $\rightarrow$
  (a) redo experiment with increased $\Phi_u(\omega)$ at those frequencies, or
  (b) increase number of data (length of experiment)
- Related analysis can be developed for $H(e^{i\omega}, \hat{\theta}_N)$

Minimum variance estimator

Central question in estimation theory:

Does there exist - in a specified situation - a lower bound for the variance of a parameter estimator?

Cramér-Rao lower bound (CRLB)

Consider observations from a random variable $\mathbf{y}$ with pdf $f_{\mathbf{y}}(y, \theta)$, where θ is the unknown parameter. Then for *any* unbiased estimator $\hat{\boldsymbol{\theta}}$ of the parameter θ, its covariance matrix satisfies the inequality

$$cov(\hat{\boldsymbol{\theta}}) \geq J^{-1}$$

with the Fisher Information Matrix:

$$J = \mathbb{E}\left\{ - \frac{\partial^2}{\partial \theta^2} \log f_{\mathbf{y}}(\mathbf{y}; \theta) \Big|_{\theta=\theta_0} \right\}$$

Remarks on CRLB

- CRLB requires knowledge of pdf $f_{\mathbf{y}}(\mathbf{y};\theta)$
- CRLB generally requires exact knowledge of θ_0
  Exception: Gaussian pdf's with linear regression model
- It provides a lower bound for unbiased estimators
- Independent of the particular estimation method
- Useful for issues as
  - analysis, and
  - experiment design
- Estimator that reaches CRLB does not necessarily exist!

Property of Prediction Error (PE) method

If in the PE setting, the noise disturbance e is Gaussian distributed and $\theta^* = \theta_0$, then the PE estimate has Maximum Likelihood (ML) properties

Properties of the ML estimator

for $N \to \infty$:

$$\sqrt{N}(\hat{\theta}_N - \theta_0) \to \mathcal{N}(0, NJ_N^{-1})$$

with J_N the Fisher Information Matrix, so that asymptotically in N

$$cov(\hat{\theta}_N) = J_N^{-1} \quad \text{(Cramér-Rao lower bound)}.$$

This implies that the ML estimator

- is consistent, provided that u is p.e. of sufficient order, and $\mathcal{S} \in \mathcal{M}$;
- asymptotically reaches the smallest possible variance (CRLB) over all unbiased estimators

no guarantees for properties in case of finite N

Conclusion

> The minimum possible variance of $\hat{\theta}_N$ is reached (asymptotically) if we estimate under the condition that $\theta^* = \theta_0$, i.e. we model both G_0 and H_0 accurately

This is a key reason for modelling H_0.

Summary

- Parameter variance decreases with signal-to-noise ratio, and with longer data records (N).
- Parameter uncertainty intervals (for a fixed probability level) can be described as ellipsoidal regions in parameter space.
- There is an analytical expression for the asymptotic covariance matrix, that can be estimated from data.
- Parameter uncertainty intervals can be converted to uncertainty intervals on the model's frequency response.
- The Cramer-Rao lower bound (CRLB) is the smallest variance that is possible for any unbiased estimator.
- Asymptotically, PE methods in the case $\theta^* = \theta_0$, have the maximum likelihood property, and reach the CRLB.